Report of the
Astrophysics Archives Program Review
For the Astrophysics Division, Science Mission Directorate
6-8 May 2015

Review Panelists: Stephen Kent, Chryssa Kouveliotou, David Meyer, H. Richard Miller (Chair), David Schade, James Schombert, Alexander Szalay, Suresh SanthanaVannan

Introduction

The Astrophysics Archives Programmatic Review (AAPR) replaces the Astrophysics Archives Senior Review, which was last held in 2011. NASA's Astrophysics Division, Science Mission Directorate regularly requests comparative reviews of Archival Data Centers with the goal of maximizing the scientific return of these centers in the present constrained funding environment. The AAPR will normally be held every three to four years to conduct an independent, comparative evaluation of the activities of NASA's Astrophysics Archives and their associated data centers. The AAPR will evaluate proposals for continued and augmented funding for a number of significant and important projects.

The scientific value and importance of user-friendly, easily accessible data archives and the expanding role of archival centers was clearly recognized in "New Worlds, New Horizons", the 2010 decadal survey report from the National Academy of Science.

The purpose of this comparative review is to assist NASA in maximizing the scientific productivity enabled by the Astrophysics Archival Program. NASA will use the assessments from the AAPR to refine its implementation strategy for the archives. They will also be used to achieve astrophysics strategic objectives and meet community requirements, and to prioritize tasks and activities for and within individual archive centers.

A new feature of the present Programmatic Review is the inclusion of a progress review of the management and maintenance of the infrastructure of the newly re-structured Virtual Astronomical Observatory (VAO). The coordinated effort by the NASA archives to manage and maintain the VAO infrastructure will be referred to as the NASA Astronomical Virtual Observatories (NAVO). However, NAVO is not part of the comparative review.

The NASA Archival Centers which are reviewed here are:

- Mikulski Archive for Space Telescopes (MAST)
- Infrared Processing and Analysis Center (IPAC)
  - Infrared Science Archive (IRSA)
  - NASA Extragalactic Database (NED)
- NASA Exoplanet Archive (NEA)
- Astrophysics Data System (ADS)
- High Energy Astrophysics Science Archive Research Center (HEASARC)

The Charter and Purpose of the Review

NASA's Astrophysics Division, Science Mission Directorate charters the Astrophysics Archival Programmatic Review panel to review the future plans presented in the proposals and assess the extent to which the Archives (not in priority order):

1) *Use their unique scientific and information technology skills to work with Astrophysics flight projects during all phases of a mission investigation to aid project personnel to design datasets*

and data production systems, and to identify and restore datasets from past Astrophysics missions to aid in current research;

2) *Ingest new datasets, and newly derived datasets created by mission teams or other NASA-funded investigators, into a curated archive;*
3) *Promote and enhance the use of NASA Astrophysics data by the scientific community and the public, by supporting research employing the archived data, analysis software, and specialized search and visualization capabilities to enhance discoverability. Feedback from user groups, and usage metrics for services, may be valuable indicators of performance;*
4) *Ensure that the data are secured against natural disasters and unauthorized access and modification;*
5) *Use current techniques in information technology in order to improve the data labeling and formatting standards that are essential to archive accessibility and improved cost effectiveness;*
6) *Where plans include archiving non-NASA funded data or data products, assess the degree to which this activity enhances the scientific return of NASA Astrophysics flight missions; and,*
7) *Demonstrate efficient management of resources, including personnel and budget, to provide continuity of a service that enables optimal scientific return.*

NASA's Astrophysics Division, Science Mission Directorate also charters the Astrophysics Archival Programmatic Review panel to assess the level to which the VAO Centers (not in priority order):

1) *demonstrate effective use of VO protocols to provide access to spectral and image data;*
2) *enable VO queries for all astronomical tables;*
3) *meet their proposed milestones with respect to the standardization of VO data access protocols;*
4) *demonstrate that VO activities have supported efficient access to large datasets;*
5) *demonstrate that standardized descriptions of all resources through the VO registry are still available to the community and there has been significant process in the adoption of common standards between the 3 Centers; and,*
6) *demonstrate how successful the 3 Centers have been in working with each other to improve multi-wavelength access.*

The Final Report is provided to Dr. Hashima Hasan, Program Scientist and to Dr. Paul Hertz, Director, Astrophysics Division, Science Mission Directorate.

Review Procedure

Each of the archive centers above was instructed by NASA to prepare proposals for continued and augmented funding for the period FY2016-2020 and given guidelines for content and budget presentation. Each proposal described the centers' current status, including its holdings, services and tools provided, metrics on usage, scientific contributions, and relation to NASA strategic goals, objectives and research focus. Proposals also presented descriptions of current projects and activities, as well as plans or possibilities for future development over the next 5 years. Budgets and FTE requirements were presented for both in-guide and over-guide budget requests.

The review was held May 6-8, 2015 in Washington, D.C. The review began with an overview by Dr. Paul Hertz, Director of the Astrophysics Division, outlining the current Astrophysics Division portfolio and budget, and highlighting recent changes in the status of NASA missions relevant to the Astrophysics Archives activities. Dr. Hashima Hasan, Program Scientist for the Archival Review, followed with an in-depth review of the charge to the Astrophysics Archival Program Review panel, a discussion of the budget, a brief review of the highlights of the proposals under review, and some clarifying information on the NASA Virtual Astronomical Observatory (NAVO) activities of relevance to the Astrophysics Archives.

Each of the archival centers gave presentations to the AAPR panel over the first two days of the meeting. Each center was represented by 3-4 people, who met for a 90-minute period scheduled with the panel. Prepared presentations were given which addressed the highlights of each proposal, provided updates since the proposal submission, when appropriate, and responded to several questions the AAPR panel had submitted prior to the face-to-face meeting. During this period, there was time for questions and discussion with the center personnel. The AAPR panel wishes to thank all center staff for their cooperation both in preparing very informative presentations and their responsiveness during the discussion.

Following each presentation, the AAPR panel met in a brief executive session (including NASA personnel), to discuss the presentation and identify any further questions it wished to ask of the center personnel. In several cases, the panel asked to talk again briefly with center staff to clarify points or to ask additional questions. Following the presentations from all centers on the second day of the meeting, the AAPR panel discussed each of the proposals in turn, identifying both its strengths and weaknesses, and reviewing its over-guide budget requests. At the end of the day, a preliminary scoring was made.

On the final day of the meeting, a final ranking was adopted and a set of recommendations was made, including priorities on over-guide requests. The panel also discussed a number of issues that were applicable to all of the archive centers. Those of note are called out below. Throughout the meeting, NASA officials were helpful in providing background information and guidance on the process of the review and were very responsive to questions from the panel.

Outcome of the Review

The final ranking of the six centers took into consideration all the criteria established in the charge to the AAPR, in both an absolute and relative sense. The panel found the first three centers (ADS, NED, MAST) to be at essentially the same, excellent level of performance and merit, and so ranked them identically. These were very strong proposals that fully responded to the AAPR Call, and contained no major weaknesses. The next two centers (IRSA and HEASARC) were performing very well but the panel identified some weaknesses in their proposals that led them to rank them slightly lower, but found little distinction between the two in overall level. Finally, the last of the centers, NEA, was found by the panel to be weaker than the others; although significantly improved compared to its predecessor, NASA Star and Exoplanet Database (NStED), which was terminated after the 2011 Astrophysics Archives Senior Review. Our evaluations and recommendations for the individual centers follow in individual sections below, in the order listed below.

- ADS – EXCELLENT
- MAST – EXCELLENT
- NED – EXCELLENT
- HEASARC – VERY GOOD
- IRSA – VERY GOOD
- NEA – GOOD

There was a serious concern expressed by the panel that the data centers' infrastructure was not up to the standards that would be expected from leading-edge data management organizations. The infrastructure and the technological approaches that are being used will certainly be obsolete at the end of the next 4-5 year review cycle. Network bandwidths available to the data centers will soon be two generations behind the current standard for research internet. The panel was disappointed at the absence of plans for major infrastructure refreshes from any of the data centers. Flat budgets create challenges to such planning and innovation but the data centers need to raise concerns about sustainability where they exist, regardless of budgetary constraints.

The panel looks to the NAVO group as being representative of the most innovative and progressive elements of the NASA data centers. We urge NAVO and other members of the data center staff to have discussions and to develop plans related to updating the infrastructure at NASA data centers with a view to bringing them back up to leading-edge standards for data management organizations.

COMMENTS ON INDIVIDUAL ARCHIVES:

ADS

The Astrophysics Data System (ADS) provides a digital library portal for researchers in Astronomy and Physics. The ADS maintains three bibliographic databases containing more than 11.4 million records covering publications in Astronomy and Astrophysics, Physics, and the arXiv e-prints. ADS also provides 77 million citation links, with

nearly half a million links to data products, SIMBAD, and NED objects. Abstracts and full-text of major astronomy and physics publications are indexed and searchable. A set of browse and visualization interfaces are also available through the user interface. In addition to maintaining its bibliographic database, the ADS also tracks citations and usage of its records to provide advanced discovery and evaluation capabilities. ADS also provides a customized interface, through the myADS user profile system for researchers to set individual search and notification preferences.

Relevancy: The ADS does not archive or distribute NASA mission data, and in that aspect it is different from the other archive centers. ADS ingest a vast physics and astronomy literature including all NASA mission relevant publications and provides links between publications, authors, and data sets. In this role, ADS provides a highly relevant service to the NASA astrophysics programs by linking important information that contributes to all stages of scientific inquiry: preparation of proposals that lead to data acquisition and/or analysis in support of NASA-related science investigations and publication and dissemination of results.

Recommendations: The panel in general agrees with the ADS prioritized list of tasks and provides the following additional guidance: (1) Maintain continuous current services, (2) Complete transition to new system including transition to the new Ingest Pipeline and incorporate functionalities from the ADS classic to the new system, (3) Improve ADS services incorporating the new database, search/indexing engine, etc., release the new user-interface, incorporate additional functionalities such as the visualization interface, and links to social interfaces such as ORCID needs to be explored.

The panel recommends that the ADS take the lead and coordinate the following activities amongst all the data archives: (1) provide tools and infrastructure (with MAST) for creating and registering digital object identifiers (DOI), and (2) work with the journals to provide direct linkages to data sets from manuscripts. The panel recommends that ADS sets up a user group, comprised of a representative user community including a member of the NASA archive community that provides guidance to ADS on (1) annual operations/development plans, (2) prioritization of new tools and infrastructure improvements, (3) applicability to science and (4) access to data.

The size of the ADS team is considerably smaller than those of other comparable bibliographic services, yet ADS provides a greater level of functionality and access to a considerably larger body of literature than the others. Going forward, ADS has identified a need to replace its current infrastructure, ADS Classic (which has grown organically over 20 years), with a new search system. However, given the current thin staffing, it is unlikely that this transition can be made in a flat funding scenario without placing current services at risk of disruption or failure. For this reason, the panel recommends that NASA fund the requested over-guide to transition to the new system and to upgrade the operational backbone.

MAST

The role of MAST is to curate and serve data from the space-borne UV/optical missions spanning nearly half a century, and to provide ancillary information and tools that will enhance the scientific impact of the NASA archive. Based at Space Telescope Science Institute (STScI), MAST is surrounded by a rich technical and academic environment with deep resources needed to maintain and support this mission.

Strengths: MAST has done an outstanding job of assimilating and storing large quantities of data from a large number of missions, which recorded a significant number of optical/UV observations. It is critical to have this repository for the astronomy community. MAST is tightly integrated with the HST project. It has also been restoring/reformatting data from past missions such as HUT. It is successfully ingesting and serving data from Kepler, even though Kepler is somewhat distinct from MAST's core expertise. In total, its holdings include 19 different missions, and more than 300TB of data. MAST is continuously ingesting data from 4 current missions (Hubble, Kepler/K2, Swift-UVOT, and XMM-OM). Data download just from the Kepler mission has exceeded 250TB. MAST has demonstrated extensive experience identifying, restoring and (in many cases) improving datasets from past missions. The site has generated phenomenal traffic, their load is more than 2 million queries

monthly. Data access is seamless. The usage of archival data yields as many scientific papers as the original PI data sets. Furthermore, the impact of the MAST archival data is very high, accounting for more than 50% of the related citations in the literature.

MAST has made substantial improvements to the infrastructure required to ingest, process, and distribute new and newly derived datasets, and are incorporating the Common Archive Observation Model (CAOM) which should permit higher levels of interoperability between various NASA and non-NASA archives. The underlying infra-structure is also becoming increasingly shared, and uses various Virtual Observatory (VO) components wherever possible. MAST is willing to use tools built elsewhere, and is open to collaborations. For example, CAOM is an underrated advancement that could very well form the core of a future VO infrastructure. MAST is applauded for working with the Canadian Astronomy Data Center (CADC) on this initiative. MAST has also benefited from the proximity of Johns Hopkins University (JHU) to implement advanced database functionality. There is a unified MAST portal that gives access to all the missions through a common interface. In addition, this is well-integrated with Virtual Observatory resources. It is easy to retrieve data, download mission documentation and analysis software, seek answers to analysis/archive questions, and cross-correlate the MAST holdings with other NASA data centers or with user-provided catalogs.

MAST has built many new value-added, innovative data products. Improvements have been made post-mission on the Galaxy Evolution Explorer (GALEX) and original Kepler mission data archives. MAST clearly enhances the usefulness of this data simply by providing a means of storing it in a common format in this database. MAST also develops higher level science products for ease-of-use and fused data source datasets. Such high level products are giving a unique flavor to MAST. Exploratory science starts with these kinds of databases. From the many such data products, we would like to highlight two here:

- The *Hubble Source Catalog* (HSC) created an object catalog from all Hubble exposures and visits, keeping an accurate track of the exposure geometries and the effective depth (exposure time). The object detections form the individual exposures have been cross-correlated to derive a rich multicolor object catalog that contains all objects that Hubble has ever looked at, in an easily searchable form. Creating such lists even for a small fraction of the Hubble fields used to be a multi-month effort. Now this data can be retrieved in minutes. Given the complexities of the HST observing strategies, there are huge differences between different fields and visits (single exposures vs. the Ultra-Deep Field), and this created significant challenges which have been successfully overcome.
- The gPhoton database is built out of more than 1 trillion photons from the GALEX mission. Each photon event is recorded as a separate entry into the database, enabling a lot of novel analyses. There is no other photon database anywhere close to this scale, although in the past, there was a lot of discussion of such databases from X-ray surveys, e.g. the ROentgen SATellite (ROSAT) photon catalog, which has never been made public.

MAST has decided to host the final archive of the Pan-Starrs project, which will exceed 2 petabytes. This was entirely funded out of the Director's Discretionary fund, and members of the scientific staff using their research time carried out the data reductions. The project to date has not added cost to the MAST budget. In turn, it brought new Big Data technologies to STScI, and as a result, they are entering the next missions which will generate data in the 100TB to1 PB range (Transiting Exoplanet Survey Satellite (TESS), James Webb Space Telescope (JWST), Wide Field Infrared Space Telescope (WFIRST), Gaia) with a well-tested set of tools.

MAST has developed an extensive range of tools for use by the community, and improved upon existing data to enhance their ease-of-use (example is Hubble Legacy Archive). Increases in distribution and publication metrics testifies to the widespread use of MAST products and services, driven by substantial improvements to MAST search and discovery tools. MAST is also developing journal linking and semantic tagging, and is also simplifying the approach to large data base queries. MAST also has extensive activities to engage user groups and collect community input for year-on-year planning.

In summary, MAST presents a clear and concrete plan for the future. It identifies key data collections and key data products as well as data services that it plans to host, create, and implement. MAST is the archive with the most innovation and a vision for the future, and they are demonstrably not satisfied just doing more of the same. Lastly,

MAST has clear metrics of success. Publications and citations are very healthy, and form a very persuasive indicator of success.

Weakness: MAST's current network speed of 100Mbits/sec is shared with all of the Space Telescope Science Institute and is grossly inadequate. The situation will become much worse in the future when new and larger data sets are ingested into the archive. They should upgrade to 10G as soon as possible.

Recommendations: The panel congratulates MAST on the many successes identified above. However, the need to upgrade the network speed to 10 Gbits/sec is critical and should be addressed as soon as possible. MAST identified a NSF proposal as a possible source of funding to support this upgrade. The panel recommends that MAST not rely on this source of funding to fulfill its core mission. The panel found that funding lines were not clear, and it was difficult to see what individual budget lines actually buy, but still the overall result is impressive. The panel recommends that MAST provide an explicit list of prioritized archive objectives, as requested in the Call for Proposals.

NED

NASA/IPAC Extragalactic Database **(**NED) is NASA's premier center for the service of extragalactic metadata, and a clearinghouse for photometric, imaging and spectrophotometric data. NED's dataset is multi-wavelength, providing a user not only with critical enabling science resources, but also with the ability to do realtime pure science within their archive using online tools through NED's website. NED provides meta-data links to data and resources at the other NASA Data Centers as well as links to the literature through ADS; thus providing a comprehensive tool for the typical user to go from data analysis to publication. NED differs from the other data centers by offering two avenues for research: (1) a central clearinghouse for galaxy data (distances, coordinates, photometry, references) and (2) discovery tools to perform analysis of multi-wavelength datasets. Most extragalactic researchers use NED as a portal to MAST and IRSA, producing a model for cross-archive access. The panel was very impressed by the range of datasets and smooth operations NED maintains for their community. The critical functions for NED are maintain a (1) comprehensive census of extragalactic objects and fundamental parameters, (2) synthesize extragalactic data and derived parameters and (3) provide services enabling multi-wavelength research. All three are well justified in their in-guide funding and NED has a solid operating plan to achieve these goals.

Strengths: NED is currently used for a large and very diverse set of research programs and continues to ingest large amounts of data (tables, images, photometry, spectroscopy) at an efficient rate. NED is essentially a cached version of a cross-match search run across nearly every data source, including many small datasets which cannot be accessed automatically (e.g., those in peer-reviewed articles). The query rate is the highest of any data archive (excluding ADS). NED has been adding new scalable tools, including support for long running queries. One of their goals has been, not just to increase the size of their holdings, but to improve its quality. They have done an admirable job. Recently NED has developed a probabilistic crossmatch tool, which is now widely deployed. NED expects to increase its data holdings by a factor of 100 over the next 5 years.

NED represents the quintessential avenue for NASA data exploitation. It is a paradigm to be followed in many areas multiple wavelength research. NED also represents the best example of cross archive access, although only for extragalactic data products. NED structure provides a model for future VO-like centers. Non-NASA datasets are well justified as coordinated observations to primary NASA datasets. NED is working well with NAVO for low level standards and ADS for high level literature metadata. The panel was particularly excited by the "Journals of the Future" program and its plans to improve the ingestion of data from published articles.

NED has an active users group and solicits input, via surveys, which have been used to help set priorities and provide NED with a basis for solid strategic planning. In past reviews, NED was criticized for not assigning science value to new datasets, thus making priority assignments impossible. This has been addressed with the

new MatchEx tool, a user survey, and continued input from the NED Users Group. Basically, NED has introduced metrics which measure which datasets are being used and now refines priorities based on those metrics. NED makes use of Infrared Processing and Archive Center (IPAC) and the IPAC Support Group (ISG) to provide security and off-site backups. The panel deemed this more than sufficient to meet security needs for the near future. NED sorely needs some sort of visiting scientist program. There are large numbers of grad students and postdocs who work only on archival data. Their data mining skills could bring to NED new ideas and, in return, provide a chance for some career starting experience at Caltech. Even a program to provide publication costs or travel to workshops (or better, a regular NED workshop at IPAC) would go a long way to improving NEDs connection to the community. NED should also consider establishing closer collaborations with small projects.

Weaknesses: The new web interface is disappointing as it is simply a re-tooled HTML version of the old interface. Style is not a substitute for a new design with better capabilities. Plans for developing web-based APIs for scripted data access are presented with little detail, and therefore it is difficult to determine its future impact. Power users are poorly served and the new interface should seek to increase the capability of power users**.**
A few large datasets dominate NED (e.g., Sloan Digital Sky Survey (SDSS) Data Release (DR) 12, not yet ingested) yet has produced over half the total redshifts (most of which are in NED already) and, oddly, these few datasets have as many galaxies as are in the entire NED holdings (although WISE will compliment SDSS in size). Thus, in round numbers, NED is simply replicating the SDSS database with respect to these galaxies. The real value is in the smaller number of objects with additional non-SDSS data, and the power of cross-archive access to non-SDSS parameters.

Physics within Galaxies was highly recommended by the previous review, but selected for cuts by the NED User's Group. The mixed signals here make the program less desirable. Progress toward ingestion of SN, HII regions is notable, NED should consider being a clearinghouse for CMD's as well.

The proposed increase in their data holdings by a factor of 100 over the next 5 years places a severe strain on the current hardware.

Recommendations: The panel recommends that the over-guide budget requested to ingest data from surveys that are still in progress might be better delayed until those surveys produce final catalogs rather than having to repeatedly ingest new versions that supersede previous versions.

No other archive asked for equipment as an over-guide request. While a demonstrated critical functionality, it does not contain an innovative component and lacks weight as an over-guide request. The proposal lacks a plan if the scale up will require more effort than 0.5 FTE. The panel recommends that this contingency should be maintained in the in-guide plan.

The Physics within Galaxies program and the Machine Learning experiment were considered highly meritorious by the panel and we support the over-guide request for these programs at a reduced level. This over-guide comes with the caveat that if critical hardware needs override the science needs, that new hardware purchases should come from this over-guide request.

The panel found that the visual and statistical database was considered too ill-defined and does not recommend funding at this time.

HEASARC

HEASARC comprises the original High Energy Astrophysics (HEA) NASA mission archive established 25 years ago, and the recently added (2002) Legacy Archive for Microwave Background Data Analysis (LAMBDA). The HEASARC team continues to provide excellent service to the community. To date the archive serves data from 39 archival missions, while planning to ingest observations from another 19 operating missions in the next 4 years. It is the dominant provider of archival data and data analysis tools for the HEA and CMB communities.

Strengths: HEASARC continues to be extremely successful in supporting the HEA user community. Their extensive expertise has been very effectively implemented in a policy of data acquisition and ingestion that starts with their involvement with a mission at its planning phase and follows it up to its legacy phase, thus guaranteeing the consistency and uniformity of the data products. This policy ensures a smooth and timely data archiving and distribution pipeline. Currently, HEASARC provides continued support of operating missions (Chandra, Fermi, Nuclear Spectroscopic Telescope Array (NuSTAR), Suzaku, Swift, XMM-Newton, and INTErnational Gamma-Ray Astrophysics Laboratory (INTEGRAL)) and plans to ingest data in the near future (starting 2017) of Astro-H and NICER. In addition, LAMBDA provides Cosmic Background Explorer (COBE) and Corot data, with a subset of Planck data, and plans to ingest data from a number of ground based observatories and Balloon experiments. The primary HEASARC data are protected by a Redundant Array of Independent Disks (RAID) system, while there is a mirror of the data inside Goddard Space Flight Center (GSFC), in case of a major disaster.

HEASARC thus supports currently the largest number of archival and current missions. The successful incorporation of LAMBDA demonstrates the ability of HEASARC to collate and curate legacy data from an extremely diverse set of space-based and ground-based observations. At the same time the quality of their products remains constant: the standardization of the data (FITS) concept is robust, seamlessly maintained and has facilitated combined and correlated mission research for a very broad community. Their curation services include configuration management for calibration data and software releases, rich metadata development for legacy missions, and services linking data to published literature. HEASARC has successfully implemented most recommendations of the last Programmatic review committee.

With creating Hera, HEASARC moved beyond providing mission data products along with processing software to be used for science analysis. With its scripting enabling capability, Hera could become a major asset of the archive for cross cutting data analysis across missions. *We recommend that HEASARC actively advertise this new capability to the user community.* As a larger user involvement requires more resources, we also recommend that HEASARC strategically plan their resource availability for the next 4 years. HEASARC's impact on the science community is evident through literature citations (significant up to 10 years after cessation of mission operations), accepted proposals using HEASARC mission data, and usage statistics. The number of papers using HEASARC mission data is impressively steady at ~ 2000 per year. Feedback is collected from advisory panels, web feedback tools, and helpdesk queries.

Inclusion of the non-NASA ground and mission data is instrumental in complementing the HEASARC archive. The main non-NASA datasets are ground and balloon-based CMB experiments and various imaging surveys in the optical and radio. The former are generally of small size, while the latter add value by providing a more complete multi-wavelength set of data for Skyview. Their implementation is at low cost and provides a unique resource for broadband research, validates the results, and allows for best exploitation of the NASA data. At the same time there are strong links to the NAVO, and the HEASARC services are International Virtual Observatory Alliance (IVOA) compliant. The interactive SkyView nicely demonstrates building a tool on top of Virtual Observatory standards.

The panel finds that the HEASARC management has very efficiently incorporated LAMBDA in the Archive showing a very effective organizational structure.

The HEASARC budget requests are covered entirely within their in-guide plan. HEASARC is currently considering the economics of outsourcing its computing needs to commercial "cloud" vendors.

Weaknesses: The panel recognizes the need for a mirror archive outside GSFC for the entire HEASARC data volume. The flat budget of the archive, and the lack of an over-guide request will result in the loss of 2FTE over the next 4 years. There is no strategic planning on the consequences of this loss. The Archive needs to effectively and actively involve their Users groups into their strategic planning. The Archive incorporates a very limited number of community originating software. With the flat budget and constantly changing OS and hardware platforms, a plan to transition the tools to an open source community based development is needed.

Relevancy: HEASARC continues to serve a large number of past and current NASA-supported missions and will

act as the data archive for two more missions and multiple balloon experiments in the future, thus remaining extremely relevant to NASA's goals.

Recommendations: HEASARC is an extremely valuable asset among NASA archives, very effectively serving the High Energy Astrophysics (HEA) and Cosmic Microwave Background (CMB) communities. The panel recommends support at the full in-guide budget. The panel encourages HEASARC to elevate LAMBDA at the same multifunctional level as the rest of the archive and to coordinate its function actively with IRSA to prevent unnecessary duplication and to optimize science.

IRSA

The NASA/IPAC Infrared Science Archive (IRSA) curates and serves scientific data products from the NASA infrared and sub-millimeter projects and missions. Along with many historical datasets such as IRAS and 2MASS, IRSA serves the Spitzer Heritage Archive, the WISE, NEOWISE and NASA Planck Archives. As NEOWISE continues to acquire new data, and Planck delivers updates (DR3), these will be added to IRSA. IRSA supports JWST by providing access to data and tools needed for planning and analyzing observations. Since the last review, IRSA has improved web portal access, VO protocols and a transition to Oracle DBMS, as recommended by the previous review panel. In addition, IRSA has made significant progress in ingesting user data products and tools. Collaboration with other Archives and long-term planning seem solid and the ingestion of time domain datasets is progressing at a good pace.

Relevancy: IRSA has strategic relevance and plays a key role in all elements of NASA's science plan, Physics of the Cosmos, Cosmic Origins, Exoplanets and Planetary Science. In addition, IRSA will be supporting several future IR missions from mission planning to archival analysis.

Strengths: IRSA is a physical member of IPAC, which hosts three distinct archives, IRSA, NED and the NASA Exoplanet Archive. These have very different functionalities, and each has been heavily used and highly visible. IRSA contains many different datasets, typically in the mid to far infrared, while NED hosts the world's largest extragalactic database, linking astronomical objects to various publications and to each other. IRSA benefits from the fact that projects producing the current major holdings, WISE and Spitzer, are both co-located at IPAC. It has a unique heritage in IR data archiving, pipelines, and products. This is a comprehensive (20 bandwidths) archive, and receives extensive usage from the community and has a very robust institutional heritage that should be preserved.

IRSA demonstrated a superb ability to work with flight projects at all mission phases, as well as the ability to identify and restore legacy datasets. In addition to fully operational mission support, the IPAC group also supports observations from instruments that have been repurposed from their original mission (Spitzer warm phase, NEOWISE). IPAC is extensively involved in JWST pre-launch planning by providing data and tools needed for planning. IPAC supports various missions by reusing the various standards, webpages, and software stacks to maximize the science reach while simultaneously reducing mission cost and implementation timeline. The reuse of the various components also helps reduce the barriers to data usage and reduces the learning curve of the various data tools by the user community.

IRSA supports several major missions, among others Spitzer, Herschel, WISE and most recently Planck. Soon they will also host the Euclid data for the US community. They will also ingest new data from Spitzer, Near Earth Object Wide-field Infrared Survey Explorer (NEOWISE) program, Stratospheric Observatory for Infrared Astronomy (SOFIA) and reprocessed data sets from Planck and Herschel. They are starting to deploy time domain data as well. They have also made an explicit effort to include user created, value-added. The IPAC/IRSA group has demonstrated a capacity to incorporate new and newly derived datasets, including (1) new observations from active missions (Spitzer, NEOWISE, SOFIA), (2) reprocessing (Herschel, Planck, some non- NASA community data sets) and (3) contributed data sets such as the multi-source Cosmic Evolution Survey (COSMOS). There is a very high usage volume from the scientific community, both for research from individual missions as well as from multi-

wavelength archival proposals. As a result it receives user contributions that are then implemented in the archive, and demonstrates good user-archive synergies.

The ability to promote and enhance the use of NASA astrophysics data is a particularly strong area for IRSA. Their proposal relates specific activities back to NASA "big questions", directly linking work back to overarching science directives. The section on Science Impact links the various IRSA mission data sets to specific astrophysics research areas. Growth in core holdings, queries, downloads and cited publications indicate increased demand for products and services from an expanding archive. IRSA has developed a range of sensor-specific and cross-sensor enhanced products (ALLWISE) and supporting web portals and services. IPAC closely aligns with its user community through User Panels and direct participation in mission-specific user groups, e.g., Spitzer. The IPAC group also provided specific, detailed responses to the previous Senior Review, much of which was directed toward user support, access and analytical tools.

In addition, the following are a list of key developments that should be continued and strengthened: (1) linkages to journals, (2) use of students and young scientists, (3) community participation, (4) improved UI , (5) data and metadata standardization, (6) API based data access, (7) visualization mechanisms, (8) web-based documentation, and (9) user interface and user support

The archive supported the download of several hundred Terabytes and tens of millions of user queries. The archive also supports some high performance computational tools like Montage. The archive will also include a user workspace and some new tools like an Spectral Energy Distribution (SED) service for multi-wavelength data. They are entering the Big Data with 85 billion rows of astronomical data in their holdings, and IRSA has started experimenting with cloud- based solutions. IRSA demonstrates good data stewardship. All data and software holdings are routinely backed up, with copies stored offsite. An uninterruptible power supply mitigates the effects of sudden outages. IPAC data centers and networks are architected, both physically and logically, to protect proprietary data or sensitive information from unauthorized access. IRSA is continuously developing new techniques for data visualization, classification and characterization. Metadata improvements include services against the archived data, enriched metadata development for enhanced access to archival products, with emphasis on cross-archive interoperability (coordinated with MAST), time-ordering of data holdings for temporal analysis and fusion of data sets for multi- frequency studies of astrophysical phenomena.

The data publication rate using IRSA data is impressive and growing. It clearly demonstrates the effectiveness of the data holdings and the access for the community (a philosophy of Turning Big Data into Better Data). An IRSA User Committee and IRSA User Surveys exist to provide feedback and guidance. Where feasible, a common interface (e.g., Gator) is used to access data from multiple, independent datasets. IRSA (and IPAC as a whole) leverage Caltech's IT infrastructure to reduce costs to the archive. The scaling up to much larger data sets has gone extremely well, they should keep up the good momentum.

Weaknesses: IRSA proposed to create Spitzer Enhanced Imaging Products (SEIP) for warm mission data being added to the Spitzer Heritage Archive. While now is probably the best time to do so, since the expertise is still available, the added value (searches for rare objects, time domain research, complement to future surveys) does not seem particularly compelling. It also seems that IRSA is being asked to pick up for the shortfall in the Spitzer warm mission budget, rather than a critical archive need.

Some of the essential improvements (to facilitate cross-linking, big data studies etc.) such as metadata enhancements are included under over-guide. IRSA should consider moving this to in-guide. For example, a combination of Spitzer and Gaia will stretch the RR Lyrae map out to 10x6 kpc. But it is unclear how this combination of data will be accomplished since the Gaia archive will be at MAST, which is an Affiliated Data Center for Gaia.

The proposal states that 10% of all papers use NASA IR data, but the actual number of such papers, and the fraction that use data from IRSA, is not given.

While a passing reference is made to recommendations of the user panels, few details are provided. This appears to be the primary channel through which new products & services are identified, prioritized and improved, but the specific role of the user panels is not clear. The overarching question for a user's group is how does IRSA

"synthetize" requirements from these various sources for planning purposes? While much work is planned, there does not seem to be a clear mechanism to determine priorities. No mention is made of what feedback, if any, is provided by the User Committee. It would be easier to evaluate the budget items knowing how user community inputs affect prioritization and planning.

Besides the currently operating missions (Spitzer warm and NEOWISE), the only "new" missions in the coming years for which IRSA is the designated archive are SOFIA and a balloon experiment. The proposal contained no comment regarding any specific needs required by JWST or WFIRST that would benefit from advanced development work at IRSA.

User support is budgeted at at twice the FTE level of other archive centers for this activity. Is this due to high usage by the community, a large number of new users, or poor documentation.

Plans (e.g., Metadata Update) are not well defined and fall mainly within the Over-Guide. It was unclear whether the requested over-guide satisfies the ingestion of new data during a "lean" IR period of IR missions. High cost of labor makes the over-guide request excessive in terms of return for the data.

Recommendations: IRSA curates and serves archives that play an important role in astrophysics and offer the potential for continued and expanded discovery. The panel fully supports the IRSA in-guide request to continue and enhance these products plus the in-guide request for ingesting new data, maintain vital archive functions, and enabling cutting-edge research**.** The panel realizes that proper instrument support enhances the scientific value of the IRSA new and planned datasets, but feels that the proposal did not provide sufficient justification for the specific staffing request for science user support. The panel finds that NASA should examine these staffing and support requirements critically before providing the necessary funding.

Planck release V3 – Questions arose regarding the redundancy of the Planck archive both with ESA and within NASA itself (LAMBDA). The panel felt the proposed V3 release has substantial scientific merit, but the question arose whether this release will be necessary if ESA makes similar improvements in the near future. The panel accepts the IRSA justification that the services provided there are unique and relevant to the astrophysics community, but feel that the budget is poorly justified in a time of limited available funding. We find that a reduced over-guide request would suffice to produce a scaled-back version of this release focused on ingesting the dataset and archive functions.

The panel felt that the amount of funding supporting cloud computing may not be sufficient to demonstrate significantly improved capabilities, and that the work proposed in the over-guide should be accomplished through in-guide funding.

Spectral cube viewer – The panel felt that, although there is a significant benefit to the proposed developments, this work is not considered critical to the core function of the archive and should be deferred or funded through alternate means.

The panel felt the over-guide request for increased user support was not appropriate for something that should be considered a core function of the archive. Such user support should be considered within the baseline funding.

NASA Exoplanet Archive (NEA)

The NASA Exoplanet Archive supports research and mission planning by the professional exoplanet community by operating a service that provides data on all confirmed and Kepler-selected candidate planets, numerous projects and contributed data sets, and integrated analysis tools. The archive includes interactive tables containing properties of all published exoplanets; numerous data products from the Kepler mission, data from the CNES CoRoT mission and from several ground-based surveys; and spectra and radial velocity measurements from the literature. Tools include a transit ephemeris predictor, light curve viewing utilities, and a periodogram tool.

After the last Archive review, NASA HQ instructed the NExScI (NASA Exoplanet Science Institute, of which NEA is a part) to recenter the prior NStED project to focus on known exoplanets, including those derived by Kepler.

Relevancy: The field of exoplanets is young and growing rapidly, with a succession of observational techniques offering revolutionary insights into the demography and physical natures of planets outside the Solar System. Current missions such as Kepler/K2 and Spitzer and future NASA missions such as JWST, TESS, and WFIRST will either discover and/or characterize exoplanets in large numbers. While other archives will capture and archive the low-level data, NEA will be responsible for capturing all the higher level science data products from these disparate missions, along with complementary ground data, into a single archive.

Strengths: The NEA has successfully recentered its efforts to focus on known exoplanets (over 1800 at present) and make available higher level data products from Kepler. It has also completed other tasks, including setting up a clearinghouse (now the Community Followup Observing Program (CFOP)), which is operated separately from the archive proper, to collect follow-up observations from the community. The CFOP is important because follow-up observations add considerable value in vetting exoplanet candidates and providing additional information about their properties. NEA offers various analysis services, such as an ephemeris service for future transit timings, a light-curve plotting tool, and a periodogram tool. Finally NEA has other sources of data on exoplanets such as CoRoT and a myriad of ground-based data sets. Such ingestion is relevant because Kepler only accounts for about half of all known exoplanets to date. As a measure of its relevance, over 90 publications of all types referenced either the NEA or the paper describing it, which is not much smaller than the number that referenced NED at a comparable stage of its development. Additionally, the ability for users to download bulk data (presumably primarily light curves) has been exercised heavily.

Weaknesses: The principle weakness is that NEA operates in a competitive, mainly PI-driven field, and the number of known exoplanets is still small. The NEA is not unique in offering lists of all known exoplanets or providing services. One value-added service being provided by the NEA is "one-stop shopping" for all the underlying data sets for NEA; the impact of this value is difficult to measure. Another value-added service is providing tools for filtering and analysis, but again, it is unclear how much value they bring. Interfaces to extract all data for a particular object could be improved – e.g., there is no way to go from the summary table of known exoplanets to the detailed data sets stored for any given planet. Finally, data in the CFOP are not available without creating a user account and logging in. It is understood the reasons for this barrier to be included, was that it was a condition of the Kepler team and that the data are "uncurated" – however, propagation of the data from the CFOP into the archive should be a goal for the coming years.

Going forward, the plan calls for ingestion of data from the literature and from a number of operating experiments (including K2), development of web-enabled versions of a plethora of tools, enabling of VO support, and expansion of the CFOP to become the Exoplanet Followup Observing Program (ExoFOP). The next mission that will produce large numbers of exoplanet candidates is TESS, yet, the role that the NEA will play in this mission or the amount of development work needed in advance is only briefly described. The largest item in the budget is for development of new software and analysis tools; it is unclear to what extent these tools are needed, since the main work seems to be web-enabling tools developed by others. The one unique service offered by the NEA is the CFOP/ExoFOP; however, support for a person to manage the activities of the external community is placed in the over-guide budget.

Recommendation: The NEA has done a commendable job re-centering its activities on known exoplanets. It still seems to be seeking the right path to make itself the center of archival research in a highly competitive field. It will take time, as advances in the field happen in sporadic bursts. The panel recommends continuing funding the NEA at the full in-guide budget but does not recommend funding the over-guide budget. The proposed over-guide work (coordination of ground-based observations) supports a priority 1 objective; it was unclear why this work was placed in the over-guide budget while other lower-priority work was included in the in-guide budget. We suggest

reconsidering whether funding of this task could be carved out of the in-guide budget by deferring work on lower-priority tasks. We additionally suggest that the NEA continue evaluating the role(s) it can best serve to increase its prominence in the exoplanet field.

NAVO

It is too early, after only a few months, to effectively evaluate progress for NAVO. Nevertheless, NAVO has created a management plan, has done an analysis of VO protocol implementation and gaps across the NASA data centers, and they report the development of a good working relationship. So things look positive at this point and progress is as expected. One of the first goals of NAVO was to realize an effective transition from the VAO era to the NAVO era. They should be congratulated on the success of this endeavor. It was not trivial.

Concerns were expressed about the history of VO development in the US where NASA data centers were involved. There were worries that there was potential for good money following bad, and it will be on the shoulders of the Project Scientist to monitor work effort across all the Data Centers.

The panel would urge NAVO to guard against being isolated. The goal of NAVO is to coordinate the efforts from the NASA data centers to realize the implementation services that are compliant with IVOA standards. The goals of NAVO must include the interoperability of NASA centers with other national and international data centers as well. NAVO should engage the US community and its data facilities and engage the world and its data facilities. NAVO must engage strongly with IVOA. They should approach IVOA work with a strategy in mind that reflects the needs of NASA and the US community while recognizing that the interests of US astronomy overlaps to a high degree with international astronomy. The work of NAVO must ensure that IVOA standards are what the communities need and what NAVO needs to implement.

Notwithstanding the engagement with IVOA in developing standards, NAVO must recognize that the core mission of NAVO is data access (including key catalog resources) for the US community through services that use VO protocols. This should be the primary focus and is expected to consume most of the efforts of NAVO. Registry is recognized as an important component but its value is in enabling data access through location of services and other functions. NAVO should implement the IVOA standards that support the most powerful data discovery and access services possible. The protocols that are now available and those available in the coming year (SIAv2, TAP) are capable of supporting a very substantial improvement in the NASA data centers capabilities. The panel believes that the roughly 8 FTEs of effort is perfectly adequate to support the achievement of strong progress on short timescales.

The panel recommends that NAVO should be reviewed annually in its first and second years to ensure that (1) it remains focused on its goals, (2) that the management structure is working, and (3) that the relationship between the data centers is healthy. NAVO should be able to demonstrate increased levels of implementation of VO protocols that fill the most important gaps and it should be able to increase growth in usage of VO-based services. Progress in fulfilling these latter goals should be clear after one year and dramatic after two years.

Each of the data archives user's groups should be asked to comment on the accomplishments and directions of the NAVO effort within that archive. The panel feels that this is a better path at the present time than forming a NAVO user's group. ADS should be the central DOI-effort coordinator for all the archives. Consult with archives.

The panel felt that this review was very early but that the initial work was indicative of very good progress. We expect to see success from NAVO in the coming years.

Evaluation Criteria:

1)Demonstrate effective use of VO protocols to provide access to spectral and image data;

Some protocols are already in place. It is too early to evaluate progress. Progress is reflected in the compilation of a directory showing where services and gaps exist now. This will be useful in demonstrating progress.

2) Enable VO queries for all astronomical tables;

Too early to evaluate progress.

3) Meet their proposed milestones with respect to the standardization of VO data access protocols;

Too early to evaluate progress.

4) Demonstrate that VO activities have supported efficient access to large datasets;

Too early to evaluate progress.

5) Demonstrate that standardized descriptions of all resources through the VO registry are still available to the community and there has been significant progress in the adoption of common standards between the 3 Centers;

The Registry functionality has been preserved during the transition to NAVO and the group should be congratulated on this achievement. A considerable number of VO services with standardized descriptions are available now.

6) Demonstrate how successful the three centers have been in working with each other to improve multi-wavelength access

A productive relationship between the three centers is essential for the success of this effort. It appears that the relationship is very positive at this time.